ASSIGNMENT WEEK 3:

5. How many cuboids are there in a 4-dimensional cube with 4 levels each? (1 Mark)

a) 625 cuboids
b) 725 cuboids
c) 125 cuboids
d) 525 cuboids

Ans:

$$Total\ number\ of\ cuboids = \prod_{i=1}^{n}(L_i + 1),$$

No of cuboids= (4+1) * (4+1) * (4+1) * (4+1) = 625

ASSIGNMENT WEEK 4:

15. An online gaming platform has 100,000 active users. During a specific month, 10,000 users become inactive. The platform identifies 20,000 users as being at risk of becoming inactive during that month. What is the hazard probability for the online gaming platform during that month?

a. 0.2
b. 0.6
c. 0.5
d. 0.25

Ans:

Hazard (probability) is the ratio of number of customers who stop between t and t+1 to the population at risk

$$h(t) = \frac{\#\ customers\ who\ stop\ at\ exactly\ time\ t}{\#\ customers\ at\ risk\ of\ stopping\ at\ time\ t}$$

Hazard Probability=10000/20000=0.5

ASSIGNMENT WEEK 6:

5. Imagine you're building a spam filter that classifies emails as spam or not spam. After testing your model, you get the following results:

- True Positives (TP): 100 emails correctly classified as spam
- False Positives (FP): 5 emails incorrectly classified as spam
- False Negatives (FN): 10 emails correctly classified as not spam but are actually spam

What is the recall of your spam filter? (2 Marks)

a) 0.812
b) 0.525
c) 0.909
d) 0.455

Ans: Recall measures the ability of a model to correctly identify all relevant instances, in this case, the proportion of actual spam emails that were correctly classified as spam.

Recall = TP / (TP + FN)

Given:

- True Positives (TP) = 100
- False Negatives (FN) = 10

Using the formula: Recall = 100 / (100 + 10) = 100 / 110 = 0.909

10. In a medical study evaluating a diagnostic test for a certain disease, 150 patients were tested. Of these, 90 patients were diagnosed with the disease, while 60 patients did not have the disease. The model predictions are as follows:

	Test Positive	Test Negative
Actual Positive	80	10
Actual Negative	20	40

Calculate the error rate of the diagnostic test.

A)0.25
B)0.2
C)0.15
D)0.18

Choose the correct option that represents the error rate of the diagnostic test based on the provided classification outcomes.

Ans: B) 0.2

error rate, misclassification rate | $\frac{FP+FN}{P+N}$

ERROR RATE = (20+10) / (90+60) = 30/150 = 0.2

ASSIGNMENT WEEK 7:

14. Consider a dataset with a binary target variable (0 or 1) and a split based on a feature resulting in two child nodes after the split.

- Node 1 (left child): Out of 40 samples, 30 belong to class 0 and 10 belong to class 1.
- Node 2 (right child): Out of 60 samples, 20 belong to class 0 and 40 belong to class 1.

which option has the correct Gini indices of the child nodes? (3 Marks)

a) Gini index for Node 1: 0.375, Gini index for Node 2: 0.444
b) Gini index for Node 1: 0.375, Gini index for Node 2: 0.320
c) Gini index for Node 1: 0.425, Gini index for Node 2: 0.320
d) Gini index for Node 1: 0.444, Gini index for Node 2: 0.375

Ans:

$$Gini\_index = 1 - \sum_{i=1}^{n} p_i^2$$

For Node 1 (left child):

- Total samples = 40
- Proportion of class 0 = 30/40=0.75
- Proportion of class 1 = 10/40=0.25

Gini_index_Node1=1−(0.75^2+0.25^2) =0.375

For Node 2 (right child):

- Total samples = 60
- Proportion of class 0 = 20/60 = 0.3333
- Proportion of class 1 = 40/60 = 0.6667

Gini_index_Node2=1−(0.3333^2+0.6667^2) = 0.4445

a) Gini index for Node 1: 0.375, Gini index for Node 2: 0.444

ASSIGNMENT WEEK 8:

14. If the true positive value is 10 and the false positive value is 15, what is the precision score for the classification model? (1 Mark)

A. 0.6

B. 0.4

C. 0.5

D. None of the above

Ans: Precision measures the accuracy of positive predictions made by the model.

Precision = True Positives / (True Positives + False Positives)

Given:

- True Positives (TP) = 10
- False Positives (FP) = 15

Using the formula: Precision = 10 / (10 + 15) = 10 / 25 = 0.4

So, the correct answer is: B. 0.4’

ASSIGNMENT WEEK 9:

11. In a 3-dimensional space represented by coordinates (x, y, z), two cluster centroids, A and B, have coordinates A(2, 4, 6) and B(5, 1, 3) respectively. What is the precise Euclidean distance between these centroids, denoting their dissimilarity in the cluster space? (1 Mark)

A) 5.20 units

B) 3.00 units

C) 4.36 units

D) 6.48 units

Ans: To find the Euclidean distance between two points in a 3-dimensional space, we use the formula:

$$\text{Distance} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2}$$

Given the coordinates:

- Point A: (2, 4, 6)
- Point B: (5, 1, 3)

Substitute the coordinates into the formula:

Distance=sqrt ((5−2)^2+(1−4)^2+(3−6)^2) =5.20 units

13. Suppose that a customer transaction table contains 9 items and 3 customers. What is the Jaccard coefficient (similarity measure for asymmetric binary variables) for C1 and C2?

	ITEM 1	ITEM 2	ITEM 3	ITEM 4	ITEM 5	ITEM 6	ITEM 7	ITEM 8	ITEM 9
C1	0	1	0	0	0	1	0	0	1
C2	0	0	1	0	0	0	0	0	1
C3	1	1	0	0	0	1	0	0	0

a. 0.75
b. 0.25
c. 0.35
d. 0.85

Ans: b. 0.25

$$J(C1, C2) = \frac{a}{a+b+c} = \frac{1}{1+1+2} = 0.25$$

15. Assume you want to cluster 7 observations into 3 clusters using the K-Means clustering algorithm. After first iteration, clusters C1, C2, C3 have following observations:

C1: {(2,2), (4,4), (6,6)}
C2: {(0,4), (4,0)}
C3: {(5,5), (9,9)}
What will be the Manhattan distance for observation (9, 9) from cluster centroid C1 in the second iteration?

Options:
A. 10
B. 5*sqrt(2)
C. 13*sqrt(2)
D. None of these

Solution: (A)

Manhattan distance between centroid C1, i.e., (4, 4) and (9, 9) = (9-4) + (9-4) = 10

ASSIGNMENT WEEK 11:

10. If a neural network has 16 input neurons and 4 output neurons, how many neurons would be recommended for the hidden layer according to thumb rule? (1 Mark)

A) 8 neurons
B) 4 neurons
C) 2 neurons
D) 12 neurons

Answer: A) 8 neurons

Thumb rule: sqrt(no of i/p neurons*no of o/p neurons)

=sqrt (16*4) = sqrt (64) = 8

ASSIGNMENT WEEK 12:

10. In a text corpus comprising 200 documents, the word "forest" and “wildlife” doesn’t co-occur in 120 documents. Both "forest" and "wildlife" co-occur in 50 documents. Furthermore, "forest" without "wildlife" appears in 10 documents, and "wildlife" without "forest" appears in 20 documents. What is the Phi coefficient to measure the correlation between the appearance of the words "forest" and "wildlife" in this dataset?

0.19
0.66
0.72
0.85

	Has word Y	No word Y	Total
Has word X	n_{11}	n_{10}	$n_{1\cdot}$
No word X	n_{01}	n_{00}	$n_{0\cdot}$
Total	$n_{\cdot 1}$	$n_{\cdot 0}$	n

$$\phi = \frac{n_{11}n_{00} - n_{10}n_{01}}{\sqrt{n_{1\cdot}n_{0\cdot}n_{\cdot 0}n_{\cdot 1}}}$$

How much more likely it is that either **both** word X and Y appear, or **neither** do, than that one appears without the other (-1 to +1)

	WILDLIFE	NO WILDLIFE	
FOREST	50	10	60
NO FOREST	20	120	140
	70	130	
	PHI	0.663388066	

PHI= (50*120-10*20)/SQRT (60*140*70*130)= 0.663